

Manuscript version: Author's Accepted Manuscript

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:

<http://wrap.warwick.ac.uk/116857>

How to cite:

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

Data mining application in assessment of weather-based influent scenarios for a WWTP: Getting the most out of plant historical data

Sina Borzooei^{1*}, Ramesh Teegavarapu², Soroush Abolfathi³, Youri Amerlinck⁴,
Ingmar Nopens⁴, Maria Chiara Zanetti¹

1. Department of Environment, land and infrastructure Engineering (DIATI), Politecnico di Torino, Torino, Italy.
2. Department of Civil, Environmental and Geomatics Engineering, Florida Atlantic University Boca Raton, USA.
3. Warwick Water Research Group, School of Engineering, The University of Warwick Coventry, UK.
4. Department of Data Analysis and Mathematical Modelling, Faculty of Bioscience Engineering, Ghent University, Ghent, Belgium.

*Corresponding author: sina.borzooei@polito.it

This is the pre-print version (before peer- review) of manuscript. The post-print version (after peer- review) can be found online:

Borzooei, S., Teegavarapu, R., Abolfathi, S. et al. Water Air Soil Pollut (2019) 230: 5. <https://doi.org/10.1007/s11270-018-4053-1>

The request can be sent to authors to receive the complete version of the manuscript.

Abstract

Since the introduction of environmental legislations and directives, the impact of combined sewer overflows (CSO) on receiving water bodies has become a priority concern in water and wastewater treatment industry. Time-consuming and expensive local sampling and monitoring campaigns are usually carried out to estimate the characteristic flow and pollutant concentrations of CSO water. This study focuses on estimating the frequency and duration of wet-weather events and their impacts on influent flow and wastewater characteristics of the largest Italian wastewater treatment plant (WWTP) located in Castiglione Torinese. Eight years (viz. 2009-2016) of historical data in addition to arithmetic mean daily precipitation rates (P_I) of the plant catchment area, are elaborated. Relationships between P_I and volumetric influent flow rate (Q_{in}), chemical oxygen demand (COD), ammonium (N-NH₄) and total suspended solids (TSS) are investigated. A time series data mining (TSDM) method is implemented with MATLAB computing package for segmentation of time series by use of a sliding window algorithm (SWA) to partition the available records associated with wet and dry weather events. According to the TSDM results, a case-specific wet-weather definition is proposed for the Castiglione Torinese WWTP. Two significant weather-based influent scenarios are assessed by kernel density estimation. The results confirm that the method suggested within this study based on plant routinely collected data can be used for planning the emergency response and long-term preparedness for extreme climate conditions in a WWTP. Implementing the obtained results in dynamic process simulation models can improve the plant operational efficiency in managing the fluctuating loads.

Keywords

Waste water treatment plant, Combined sewer system, Data mining, wet-weather, historical data

1. Introduction

Combined sewer systems (CSSs) are designed to collect surface runoff in addition to municipal and industrial wastewater. During heavy rainfall when the volume of wastewater in CSSs exceeds the capacity of the collection system or connected treatment plant, combined sewer overflows (CSOs) discharge directly to a surface water body (Burian et al., 1999). The significant chemical, physical and biological impacts of CSOs on receiving water bodies are well documented (e.g. Field and Sullivan, 2001). The adaptation of urban water and wastewater framework directives (CEC, 1996;1991) made these untreated or partially treated wastewater streams, a priority concern (Mostert, 2003). To study the adverse impacts of CSO on the receiving water quality, attention has focused on the qualitative and quantitative analysis of wet-weather flow (WWF) and its influence on treatment plant performance (Clark et al., 2007). The quality and quantity of WWF depends on several factors including the size and layout of the sewer system, land use patterns and the impact of the urbanization on them, duration, intensity and areal extent of wet-weather events (Kothandaraman, 1972). Since a holistic approach to consider all the parameters affecting WWF is not a straightforward task, analysis of historical treatment plant data can be an alternative providing this crucial information for managing the fluctuating load during wet-weather events (Suarez and Puertas, 2005). Several studies have focused on elucidating empirical relationships between precipitation intensity (P_I), influent flowrate (Q_{in}), and wastewater characteristics (Berthouex and Fan, 1986; Giokas et al., 2002; Karagozoglu and Altin, 2003; Mines et al., 2007; Rouleau et al., 1997) and concluded that a positive correlation between P_I and Q_{in} exists. In addition, existing studies (Bertrand-Krajewski et al.,1995; McMahan, 2006; Rouleau et al.,1997; Stricker et al.,2003) have demonstrated that an increasing Q_{in} in wet-weather conditions resulted

in increased loadings of influent wastewater characteristic parameters including chemical oxygen demand (COD), TSS, BOD₅, ammonia, total Kjeldahl nitrogen (TKN) and fecal coliforms. The literature shows that the majority of existing studies focused on seasonal or monthly average P_I for investigating the impact of wet-weather on WWTP's influent parameters, while little attention has been paid to daily variability of rainfall quantities. Daily P_I was not considered in previous studies due to high incidence of zero rainfall records and non-identified minimum precipitation which can affect specific plant influent data (P_{th}) and plant upset time (t_u) after each wet-weather event (Oliveira-Esquerre et al., 2004).

In order to propose an accurate wet-weather definition for a WWTP, a robust prediction method for P_{th} and t_u values is needed. Since P_{th} and t_u parameters are case specific and highly dependent on the length and structure of sewer systems along with the plant's operating condition, extensive sampling and measurement campaigns were recommended for determining of these parameters (Berthouex and Fan, 1986).

Time Series Data Mining (TSDM) is one of the most commonly used methods for data mining problems which involve temporal variation aspects. The TSDM method has been successfully applied for understanding features of high dimensional time series datasets including similarity search, clustering, classification, segmentation and motif discovery (Fu, 2011; Antunes and Oliveria, 2011; Lovrić et al., 2014). Segmentation of time series is a pre-processing step in the analysis of the temporal sequences applied to extract internally homogenous segments for reducing the dimensionality of the data sets and discovering the patterns and rules present in behavior of observed variables (Chundi and Rosenkrantz, 2009; Chung et al., 2004; Fu et al., 2001; Gionis and Mannila, 2003). Amongst the most commonly used segmentation algorithms including Top-

Down, Sliding Window and Bottom-Up, the Sliding Window algorithm (SWA) also known as one-pass algorithm is the desirable choice due to its procedural simplicity (Lovrić et al., 2014).

In this study a quantitative analysis of the impacts of rainfall events on influent flowrate and associated water quality constituents for the Castiglione Torinese WWTP was performed by use of 8 years historical data of the plant. Time series segmentation is applied by means of the SWA methodology and the plant specific wet-weather definition is proposed accordingly. Detailed statistical analyses are conducted to compare influent loadings and flowrate under two weather-based scenarios.

2. Material and methods

2.1 Data acquisition

The Castiglione Torinese WWTP is located 11 km Northeast of Turin, the capital of Piedmont state, in the Northwest of Italy (Fig. 1). The plant is treating 590,000 m³/d of combined municipal and industrial wastewater, corresponding to an organic load of 2.1 million of equivalent inhabitants. The wastewater load consists of sewage and runoff from 38 municipal catchments in the Piedmont region which flows to the plant and is, after treatment, discharged into the River Po. There are four major wastewater treatment modules each consisting of a primary clarifier followed by biological nutrient removal (BNR) activated sludge (AS) system. From 2009 to 2016, the wastewater quality parameters were measured by 24 hours composite sampling of the plant's influent during 2920 consecutive days. Inlet N-NH₄⁺ (N-NH₄), COD and TSS concentrations were determined based on the CNR-IRSA methodology (IRSA, 1994). During the data sampling period, the influent flowrate was continuously measured with a 5-minute interval by ultrasonic flowmeters

installed at the entrance of each wastewater treatment module of the plant. Total daily flowrate of the plant (Q_{in}) was calculated by summing up the daily average flowrates of each module.

Precipitation data used in this study were provided by the Piedmont Environmental protection agency (Arpa Piemonte, 2016). Eight meteorological stations equipped with tipping bucket rain gauges were selected in the catchment area to collect the daily Precipitation data from 2009 to 2016 (Fig. 1). An arithmetic mean method was adopted to convert point precipitations at different meteorological stations into a uniform value for the whole catchment area. According to the arithmetic mean method, given the coordinates of n meteorological stations along with their respective recorded precipitation values ($P_j, j = 1, 2, \dots, n$), mean precipitation over the catchment area for a given time (P_I) can be determined from (Eq. 1)

$$P_I = (1/n) \sum_{j=1}^n P_j \quad (1)$$

The dataset has 17,892 records in total with 47 missing data points. A daily value of the stations with missing data was excluded for determining the arithmetic mean. The city of Torino has undertaken sewer separation project so logically rainfall in this area is not collected by combined sewer system; however, three meteorological stations in Torino (viz. station #2, 5, 6) were included in calculation of arithmetic mean for assessing precipitation rates of the areas close to Torino such as Borgaro Torinese.

2.2 Data preparation and treatment

All parameters in the dataset, excluding P_I , were screened to identify missing elements, detect outliers and exclude them from the dataset. The missing data included the days with problems in entering the data, sampling process or instruments. In the case of WWTP monitoring, an outlier usually occurs due to instrumentation or human error (Chandola et al., 2007). Considering the skewed distributions of data parameters including TSS and COD, application of simple quartile-

based methods such as box-and-whiskers plot would detect a large fraction of datasets as outlier (Zhu et al., 2015). To minimize the loss of data, the statistical parametric approach of generalized extreme studentized deviate (GESD) method (Rosner, 1983) was adapted to determine the outliers of each univariate data set. In contrast to other parametric outlier detection tests such as Grubbs (Grubbs, 1969) and Tietjen-Moore (Tietjen and Moore, 1972), the GSED test only requires an upper bound for the suspected number of outliers which makes it a suitable method for this study. In GESD method, given the upper bound (r_u), r_u number of tests are performed (each for an outlier) to iteratively compute R_i from Eq.2:

$$R_i = \frac{\max_i |x_i - \bar{x}|}{s} \quad (2)$$

Where \bar{x} and s are the sample mean and standard deviation respectively. The observation which minimizes $|x_i - \bar{x}|$ is removed and R_i is computed with $n-1$ observations and this process will be repeated until r_u observations will be removed. Following to above iterative computations r_u number of critical values (φ_i) are calculated from Eq.3:

$$\varphi_i = \frac{(n-i)t_{(k,n-i-1)}}{\sqrt{(n-i-1+t_{(k,n-i-1)}^2)(n-i+1)}} \quad (3)$$

with $i=1, 2, \dots, r$ and $t_{k,v}$ is the 100k percentage point from t distribution with v degree of freedom. k is calculated from Eq.4:

$$k = 1 - \frac{\alpha}{2(n-i+1)} \quad (4)$$

Where α is the significance level. Finally, the number of outliers is determined by finding the largest i such that $R_i > \varphi_i$.

Furthermore, to be able to treat different variables equally, influent flow and concentrations were converted to $Z(x)$ values according to Eq. 5.

$$Z(x) = \frac{x - \bar{x}}{s} \quad (5)$$

Where x is the original data, \bar{x} and s are respectively the average and standard deviation of the influent variables.

Following the initial data treatment, the daily, monthly averaged and annual trends of treated datasets were investigated and hydraulic flow, concentrations and loadings peaking factors were derived from the datasets. To investigate the impact of P_I on plant influent data, scatter plots of P_I versus Q_{in} , COD, TSS and N-NH₄ were made. Linear regression analyses were performed to generate the line of the best fit and the square of the correlation coefficient, R^2 , values were estimated. The R^2 , is a quantitative indicator of the proportion of the predictable dependent variable variance from the independent variable (Edwards et al., 2008). To study the statistical significance of the derived correlations, a hypothesis F-test was conducted.

2.3 Segmentation of time series

In this study, 5 time series namely P_I , Q_{in} , COD, TSS and N-NH₄ were considered as sequence of time dependent values of the observed variables arranged by chronological order in successive period of a day. Symbolically, each of the time series (T), as a set of 2728 pairs of data can be represented as follows:

$$T_v = \{(v_1, t_1), (v_2, t_2), \dots, (v_i, t_i), \dots, (v_n, t_n)\} \quad (6)$$

with $i=1, 2, \dots, n=2920$ as a number of available observations, v_i is the value of the observed variable and t_i is the time (d) on which the value was recorded. In the segmentation process, each time series is divided into series of segments as consecutive portions. Segmentation, S , represents the time series, T , in a form of a set of m consecutive segments as follows:

$$S = \{S_1, S_2, \dots, S_j, \dots, S_m\} \quad (7)$$

With $j = 1, 2, \dots, m$, each segment, S_j , consists of a certain number of pairs of data from the original time series.

Each data point can be represented by S_{ij} where i denotes the order of the point in the original time series and j denotes its corresponding segment. In each segmented series, variables belonging to that segment can be represented by a specific value (v_s) such as an average of a segment, or a function which is suited to data of the segment (Lovrić et al., 2014).

The initial step for implementing SWA is the determination of the left boundary of the first potential segment. The first data point of each time series is considered as the left anchor of the first potential segment. Considering the nature of the available time series and the partitioning objective of this study, for the pre-defined length of the segment (L_s), the second segment is created by sliding down the sequence for the single unit of the time while the size of the window remains constant (Fig. 2). This formation process of the segments repeats until each original time series is covered entirely and the stopping point of the last formed segment becomes the last point of the original time series.

The representative value (v_{sj}) for variables belonging to the S_j segment of each of the original time series were calculated and new time series were created by substituting the first pairs of data from the original time series with (v_{sj}, t_j) where t_j is the time in which the left boundary of the S_j segment was recorded in the original time series. For data partitioning purposes, the representative value (v_{sj}) of the P_I variables (P_{sj}) present in the S_j segment is defined as follows:

$$P_{sj} = \sum_{i=t_j}^{(t_j+L_s)} S_{ij} \quad (8)$$

By use of Eq.5, an accumulative precipitation was considered for the period of each segment which is the best interpretation for the continuous nature of precipitation impact on influent concentration of the treatment plant. For partitioning the observation to wet and dry weather events, the following condition was considered:

$$if P_{sj} \leq P_{th}, \{t_j \in t_{dry}\}, else \{t_j \in t_{wet}\} \quad (9)$$

where P_{th} is the pre-specified threshold value for precipitation rates. For identification of the best combination of the L_S and P_{th} parameters, 25 scenarios with different values of P_{th} (0, 1, 2, 3, 5 mm) and L_S (1, 2, 3, 5, 7 days) were developed and datasets were partitioned accordingly. Since for precipitation data, threshold exceedances are seen to occur in groups, as an extremely rainy day is likely to be followed by another, the suitable range of threshold values were selected referring to previous studies (e.g. McMahan, 2006) and engineering judgment. Assuming a linear relation between P_s and Q_{in} , the partitioning scenario with the highest coefficient of determination of positive correlation between P_s and Q_{in} of wet-weather data was selected and the wet-weather definition was proposed correspondingly. Fig. 3 provides a summary of the methods implemented in this study.

3. Discussion and analysis of results

3.1. Preliminary data treatment

The data treatment identified 112 data points (3.8% of the data) as outliers. For the flow-rate dataset no outlier was detected. Both COD and TSS data showed high incidence of missing and outlier values. The majority of COD and TSS outliers were recorded in June 2015. The plant operational data showed that in the June 2015, samples were collected from a point out of the main stream of the influent wastewater. The datasets show that there were 80 days with missing data

and 112 days with outlier values detected by GESD method, which were removed from the datasets, meaning that the remaining 2728 days were used for further study. Table 1 presents the statistical features of the datasets. It should be stressed that outlier detection process highly depend on the time scale of the data collection. For instance, using minute interval in the collection of the data will provide rather smoother data than hour or daily intervals (except pumps turning on and off).

3.2. Analyses of data and peaking factors

The data show that the monthly averaged influent flow of Castiglione Torinese has the lowest annual record, $Q_{in}=16.1 - 17 \text{ Mm}^3$, during August followed by January and December. The highest Q_{in} ($18.6 - 19 \text{ Mm}^3$) was recorded during the month of May followed by June and March. The recorded data from 2009 to 2016 shows an annual declining trend in influent flowrate (Table 2). To determine the flow peaking factors (P_{F1} and P_{F2}), the annual average daily flow (Q_{AD}), maximum monthly average daily flow (Q_{MMAD}) and peak daily flow (Q_{PD}) during each month were calculated (Table 3). Peaking factors are key parameters for flowrate design of WWTPs when historical flow data are not available. All the designed units of a WWTP must be capable of handling the peak hourly flow or in some cases Q_{PDF} , without disrupting the treatment processes. The average and range of P_{F1} value obtained in this study corresponds to the lower bound of the range reported in Metcalf et al. (1991) (1.5 – 3.0); however, a lower value was reported in Reynolds and Richards (1996).

The influent concentration data collected from 2009 to 2016 demonstrates that monthly averaged trends of COD and N-NH_4 vary significantly during a year and show peaks during winter and fall seasons. The monthly variations in the influent concentrations is due to seasonal fluctuations in water consumption in the catchment area and precipitation rates. TSS was found the least sensitive

parameter to seasonal changes. Influent COD and TSS concentration and mass loading peaking factors as key parameters for design of unit operations in WWTPs were determined. Average daily concentration (C_{AD}) and loadings (L_{AD}) were determined for parameters by averaging daily concentrations and loadings records in the studied period. Maximum monthly average daily concentration (C_{MMAD}) and loadings (L_{MMAD}) as well as peak daily concentration (C_{PD}) and loadings (L_{PD}) were estimated similarly to the Q_{MMAD} and Q_{PD} . Average and ranges of $P_{C1,3} = C_{MMAD} : C_{AD}$; $P_{C2,4} = C_{PD} : C_{AD}$; $P_{L1,3} = L_{MMAD} : L_{AD}$ and $P_{L2,4} = L_{PD} : L_{AD}$ peaking factors were calculated which are summarized in Table 3. Both peaking factors determined for COD and TSS concentrations fall within the range reported by Mines et al. (2007). Peaking factors obtained for COD and TSS loadings are consistent with the reported values in Metcalf et al. (1991).

Further, to quantitatively investigate the dispersion of recorded precipitation rates in different meteorological stations around the considered uniform P_I value, daily population (8 values per day) standard deviation of mean (SDOM) was measured (assuming statistical independence of the values in the sample) as shown in Fig. 4. The results obtained from the arithmetic mean method are accurate if meteorological stations are uniformly distributed in the catchment area and recorded precipitation rates in each individual station do not vary significantly. Fig. 4 confirms that more than 90 % of the observations show the tendency to be close to the P_I values (SDOM values between 0 and 2). Furthermore, it can be observed that data collected in days with a recorded P_I value greater than 10 mm are spread out over a wider range of values. Studying the monthly averaged P_I values during the studied period, the month of November was detected as the wettest month with $P_I = 144.3$ mm followed by April and May with 112.8 and 109.5 mm respectively. On the other hand, January, December, February were driest months with $P_I = 24.6, 40.5, 57$ mm

respectively. Fig. 5 demonstrates the variation of monthly accumulated P_I values in the studied period.

3.3. Regression analysis

The impact of P_I on Q_{in} , TSS, COD and $N-NH_4$ parameters for Castiglione Torinese WWTP were investigated by use of regression analysis (Fig. 6). The solid line in Fig. 6 shows the best fit line obtained from a linear regression method. Table 4 summarizes the results of regression analysis and statistical measures. In this study for the coefficient of determination less than 0.04 represents no or negligible correlation is considered. According to Franblau et al. (1958), coefficient of determination less than 0.04 represents no or negligible correlation. Low and moderate degree of correlation can be identified by R^2 value between 0.04 to 0.16 and 0.16 to 0.36 respectively. An R^2 value between 0.36 and 0.64 yields marked degree of correlation and $0.64 < R^2 < 1$ indicates high degree of correlation. The results presented in Table 4 and Fig. 6 show a moderate positive correlation ($R^2 = 0.25$) between P_I and Q_{in} . Increase in the rainfall coincides with a growth in infiltration and inflow into the sewer system. A low degree of negative correlation was found between P_I and the pollutant concentrations in the data which supports the dilution effect. The low degree of correlation obtained in this study are partially due to the composite sampling method, and hence loss of information about short term data behavior as reported by Schilperoort (2011), in addition to the impact of wet-weather events. In order to understand how changes in the predictors are associated with the changes in response parameters, regardless of R^2 values, the relationships should be statistically significant (Schmetterer, 2012). To investigate the statistical significance of the results, a hypothesis F-test was performed at a level of significance (alpha value) of 0.05. Under the null hypothesis of the F-test, all regression parameters are considered zero and the regression function does not depend on the explanatory variable. For rejecting the

null hypothesis, the measured F statistic (F_{st}) value should be less than F critical (F_{cr}) values obtained from statistical tables for a given level of significance. To study the overall statistical significance of the results, a P-value was also determined for each individual relation. P-values substantially smaller than critical alpha indicate that the null hypothesis can be rejected. Comparison between F_{st} and F_{cr} values (Table 4) for each test, show that the relations proposed in this study are statistically significant ($F_{st} < F_{cr}$). It was observed that all measured P values were less than alpha (0.05) which indicates the overall statistical significance of the relations obtained in this study.

3.4 Data partitioning and wet-weather definition

In total, 25 scenarios were developed to identify the best combination of the L_S and P_{th} parameters. The best correlation between P_s and Q_{in} ($R^2 = 0.35$) was obtained for the scenario with $P_{th} = 3$ mm and $L_S = 2$ days. Hence, for the Castiglione Torinese WWTP, wet-weather condition was defined as an event with accumulated precipitation rate (P_a) greater than 3 mm which occurs at least 48 hours after a previous measurable wet-weather. All data sets were partitioned to wet and dry weather data according to the wet-weather definition. From a total number of 2728 data points, 991 observations (36%) met the defined wet-weather condition and the remaining data points were classified as dry weather data.

Statistical analyses were conducted to identify the significant differences in the influent loadings and concentrations under wet and dry flow conditions. Kernel density estimation (KDE) was performed to estimate and compare the probability density functions of dry and wet sets of observations. In KDE a Gaussian kernel was used as a weighting function and the probability density functions (PDF) were assessed by computing the geometric mean of kernel function for all data sets. Given the importance of optimum bandwidth of the kernel on the accuracy of the

estimate, the normal distribution approximation method proposed by Silverman (2018) was adopted to identify the optimum bandwidths. The results presented in Fig. 7a indicate that the influent flowrates in wet-weather condition were higher by 15-25 % than those under dry weather condition. The reduction of influent TSS, COD and N-NH₄ in the wet-weather condition with dilution factors of (0.06-0.08), (0.15-0.17), (0.18-0.2) were obtained (Fig. 7 b, c and d, respectively). The inflow of wet-weather wastewater into the plant, resulted in the decreasing of all the influent concentrations due to a dilution effect. Since characteristics of organic matter represented by influent COD and concentration of nitrogen compounds such as N-NH₄ are affected by variations in concentration of household and industries, the impact of wet-weather prevails in COD and N-NH₄. The results show that TSS variation is less influenced by weather condition, only 6 - 8% variation was observed between wet and dry conditions. The accumulated pollutants in the sewer system and the impact of first-flush runoff which induce sudden increases of TSS, in the initial stage of the wet-weather events after a dry period, can diminish the dilution effect (Sansalone and Cristina, 2004).

4. Conclusions

As many largescale wastewater treatment plants, Castiglione Torinese WWTP provides treatment services over the vast range of the influent conditions by operating in a relative static mode. Accurate and well-timed measurement of influent flow and wastewater characteristics during wet-weather events can guide these large facilities toward dynamic operation to manage the fluctuating loads efficiently. However, the measurement of these variables is not a straightforward task and often involves large time delays. This study provides an assessment of two significant weather scenarios (wet and dry) in the Castiglione Torinese WWTP by use of 8 years of historical data as

well as precipitation rates recorded in 8 meteorological stations in the plant catchment area. Results from this study can be summarized as follows:

- 1) Identification of almost 4 % of the entire dataset (2920 consecutive days) as outliers in preliminary data treatment stage of this study, highlights that the data quality evaluation for historical WWTPs data, deserves a special note.
- 2) Low to moderate degree of negative correlations were observed between plant influent concentrations and arithmetic mean daily precipitation rates (P_I) of the catchment area. Higher degrees of correlations could not be achieved mainly due to the composite sampling, and hence loss of information about short term data behavior as well as the variable impact of wet-weather events.
- 3) Two weather-based clusters were classified considering daily variation of P_I time series and a case specific wet-weather definition is proposed for the plant. The wet-weather condition is defined as an event with accumulated precipitation intensity (P_a) equal or greater than 3 mm which occurs at least 48 hours following the previous measurable wet-weather event.
- 4) The data in the wet-weather cluster (36% of total historical data) indicated on average a 20 % higher flowrate, and a 6-20 % lower influent TSS, COD and N-NH₄ concentration in comparison to the dry-weather cluster.

The results obtained in this study can be implemented in process simulation models to obtain more realistic understanding about the performance, energy consumption and effluent quality of the plant in wet-weather condition.

5. Acknowledgements

The authors acknowledge SMAT (Società Metropolitana Acque Torino) financial support for this project.

References

1. Antunes, C.M., and Oliveira, A.L. (2001). Temporal data mining: An overview. In KDD Workshop on Temporal Data Mining, p. 13.
2. Arpa Piemonte. (2016). Agenzia regionale per la protezione ambientale . [ONLINE] Available at: <http://www.arpa.piemonte.gov.it/>. [Accessed 1 February 2016].
3. Berthouex, P., and Fan, R. (1986). Evaluation of treatment plant performance: causes, frequency, and duration of upsets. *J. Water Pollut. Control Fed.* 368–375.
4. Bertrand-Krajewski, J.-L., Lefebvre, M., Lefai, B., and Audic, J.-M. (1995). Flow and pollutant measurements in a combined sewer system to operate a wastewater treatment plant and its storage tank during storm events. *Water Sci. Technol.* 31, 1–12.
5. Burian, S.J., Nix, S.J., Durrans, S.R., Pitt, R.E., Fan, C.-Y., and Field, R. (1999). Historical development of wet-weather flow management. *J. Water Resour. Plan. Manag.* 125, 3–13.
6. CEC. (1996). Directive concerning integrated pollution prevention and control (96/61/EEC). *Official Journal of the European Community*, L 257, 26–40.
7. CEC. (1991). Directive concerning urban wastewater treatment (91/271/EEC). *Official Journal of the European Community*, L135, 40–52.
8. Chandola, V., Banerjee, A., and Kumar, V. (2007). Outlier detection: A survey. *ACM Comput. Surv.*
9. Chundi, P. and Rosenkrantz D. (2009) Segmentation of Time Series Data, in J. Wang (Ed.), *Encyclopaedia of Data Warehousing and Mining*, Information Science Reference, New York, USA, pp. 1753–1758.
10. Chung, L., Fu, T. C. and Luk, R. (2004) An evolutionary approach to pattern-based time series segmentation, *IEEE Transactions on Evolutionary Computation*, IEEE Press, Vol. 8, Issue 5, pp. 471-489.
11. Clark, S.E., Burian, S., Pitt, R., and Field, R. (2007). Urban wet-weather flows. *Water Environ. Res.* 79, 1166–1227.

12. Edwards, L.J., Muller, K.E., Wolfinger, R.D., Qaqish, B.F., and Schabenberger, O. (2008). An R² statistic for fixed effects in the linear mixed model. *Stat. Med.* 27, 6137–6157.
13. Field, P.R., and Sullivan, P.D. (2001). Overview of EPA's wet-weather flow research program. *Urban Water* 3, 165–169.
14. Fu, C., Chung, F. L., Ng, V. and Luk, R. (2001) Evolutionary segmentation of financial time series into sub-sequences, in: *Proceedings of the 2001 Congress on Evolutionary Computation*, Seoul, Korea, pp. 426-430.
15. Fu, T. (2011). A review on time series data mining. *Eng. Appl. Artif. Intell.* 24, 164–181.
16. Franzblau, A. N. (1958). *A primer of statistics for non-statisticians*. Oxford, England: Harcourt, Brace.
17. Giokas, D., Vlessidis, A., Angelidis, M., Tsimarakis, G.J., and Karayannis, M. (2002). Systematic analysis of the operational response of activated sludge process to variable wastewater flows. A case study. *Clean Technol. Environ. Policy* 4, 183–190.
18. Gionis, A. and Mannila, H. (2003) Finding recurrent sources in sequences, in: *Proceedings of the 7th annual international conference on research in computational molecular biology (RECOMB 2003)*, pp. 123–130.
19. Grubbs, F.E. (1969). Procedures for detecting outlying observations in samples. *Technometrics* 11, 1–21.
20. IRSA, C. (1994). *Metodi analitici per le acque*. Ist. Poligr. E Zecca Dello Stato Roma.
21. Karagozoglu, B., and Altin, A. (2003). Flow-rate and pollution characteristics of domestic wastewater. *Int. J. Environ. Pollut.* 19, 259–270.
22. Kothandaraman, V. (1972). Water quality characteristics of storm sewer discharges and combined sewer overflows (Illinois State Water Survey).
23. Lovrić, M., Milanović, M., and Stamenković, M. (2014). Algorithmic methods for segmentation of time series: An overview. *J. Contemp. Econ. Bus. Issues* 1, 31–53.
24. McMahan, E.K., (2006). Impacts of Rainfall Events on Wastewater Treatment Processes. Retrieved from <http://scholarcommons.usf.edu/etd/3846/>.
25. Metcalf, E., Eddy, H.P., and Tchobanoglous, G. (1991). *Wastewater engineering: treatment, disposal and reuse*. McGraw-Hill N. Y.
26. Mines Jr, R.O., Lackey, L.W., and Behrend, G.H. (2007). The impact of rainfall on flows and loadings at Georgia's wastewater treatment plants. *Water. Air. Soil Pollut.* 179, 135–157.
27. Mostert, E. (2003). The European water framework directive and water management research. *Phys. Chem. Earth Parts ABC* 28, 523–527.

- 28.** Oliveira-Esquerre, K.P., Seborg, D.E., Bruns, R.E., and Mori, M. (2004). Application of steady-state and dynamic modeling for the prediction of the BOD of an aerated lagoon at a pulp and paper mill: Part I. Linear approaches. *Chem. Eng. J.* *104*, 73–81.
- 29.** Reynolds, T.D., and Richards, P.A. (1996). Unit operations and processes in environmental engineering (PWS Publishing Company Boston, MA).
- 30.** Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. *Technometrics* *25*, 165–172.
- 31.** Rouleau, S., Lessard, P., and Bellefleur, D. (1997). Behaviour of a small wastewater treatment plant during rain events. *Can. J. Civ. Eng.* *24*, 790–798.
- 32.** Sansalone JJ, Cristina CM. First flush concepts for suspended and dissolved solids in small impervious watersheds. *J Environ Eng.* 2004;130(11):1301–1314.
- 33.** Schilperoort, R.P.S. (2011). Monitoring as a tool for the assessment of wastewater quality dynamics.
- 34.** Schmetterer, L. (2012). Introduction to mathematical statistics (Springer Science & Business Media).
- 35.** Silverman, B.W. (2018). Density estimation for statistics and data analysis (Routledge).
- 36.** Stricker, A.-E., Lessard, P., Héduit, A., and Chatellier, P. (2003). Observed and simulated effect of rain events on the behaviour of an activated sludge plant removing nitrogen. *J. Environ. Eng. Sci.* *2*, 429–440.
- 37.** Suarez, J., and Puertas, J. (2005). Determination of COD, BOD, and suspended solids loads during combined sewer overflow (CSO) events in some combined catchments in Spain. *Ecol. Eng.* *24*, 199–217.
- 38.** Tietjen, G.L., and Moore, R.H. (1972). Some Grubbs-type statistics for the detection of several outliers. *Technometrics* *14*, 583–597.
- 39.** Zhu, J.-J., Segovia, J., and Anderson, P.R. (2015). Defining Influent Scenarios: Application of Cluster Analysis to a Water Reclamation Plant. *J. Environ. Eng.* *141*, 4015005.

Figures

Fig. 1 Location of Castiglione Torinese plant and meteorological stations in the catchment area

Fig. 2 Schematic description of sliding window algorithm

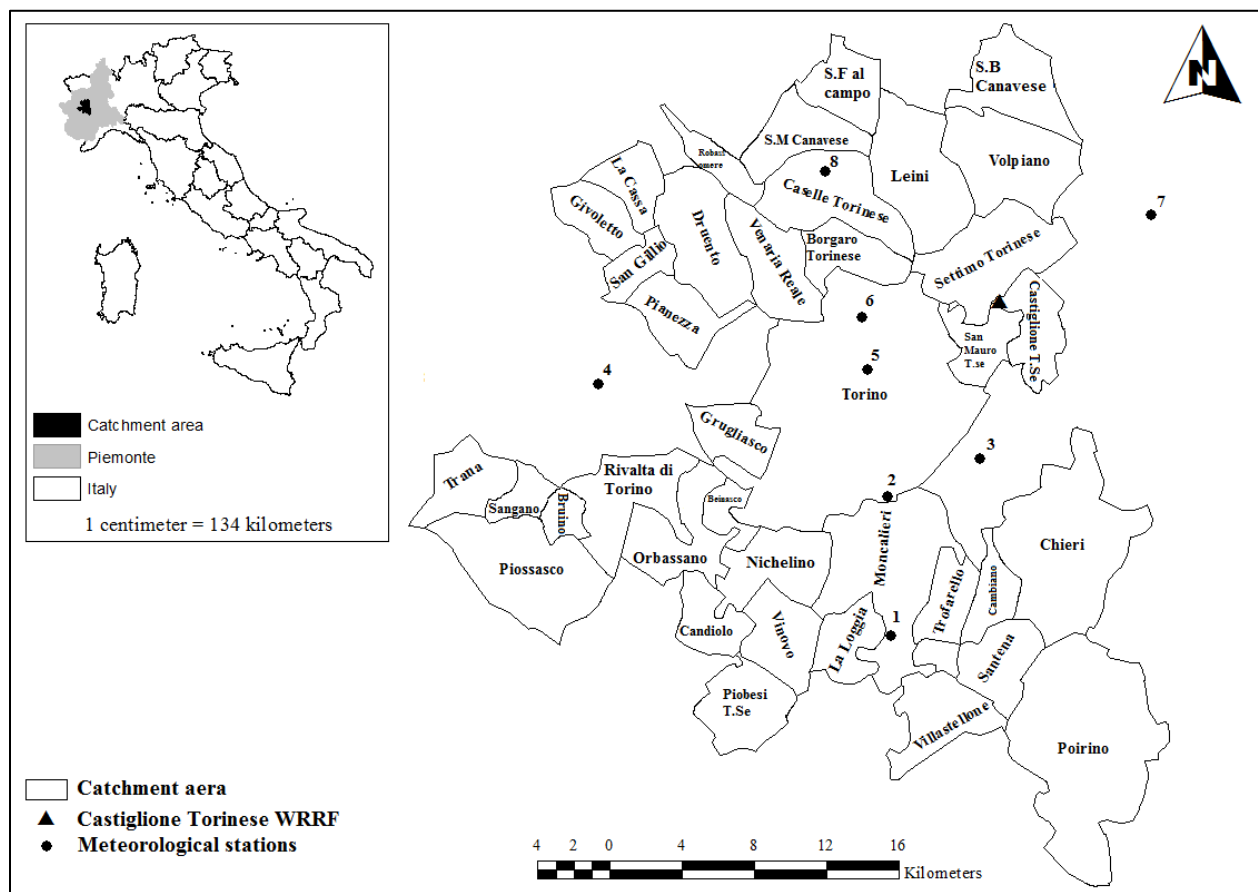
Fig. 3 A summary of the implemented methods in this study

Fig. 4 Standard deviation of mean for P_1 values

Fig. 5 Variation of monthly averaged P_1 values during the studied period

Fig. 6 Variation of Influent flow (a), TSS(b), COD(c) and N-NH₄ (d) versus precipitation intensity (P_I)

Fig. 7 Kernel density estimations wet and dry weather Influent flow (a), TSS (b), COD (c) and N-NH₄ (d)



Meteorological stations: 1) Baudduchi 2) Torino Vallere, 3) Pino T.Se 4) Rivoli la perosa, 5) Turin via the conolata, 6) Turin reiss romoli, 7) Castagneto po, 8) Casella T.Se

Fig. 1 Location of Castiglione Torinese plant and meteorological stations in the catchment area

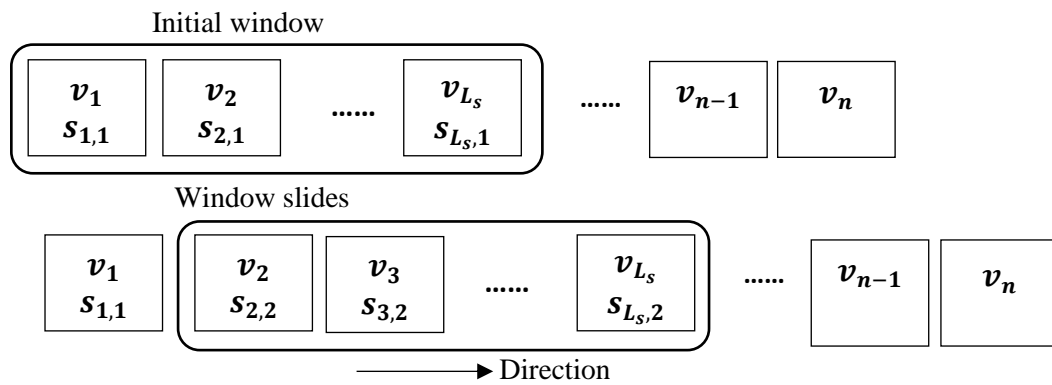


Fig. 2 Schematic description of sliding window algorithm

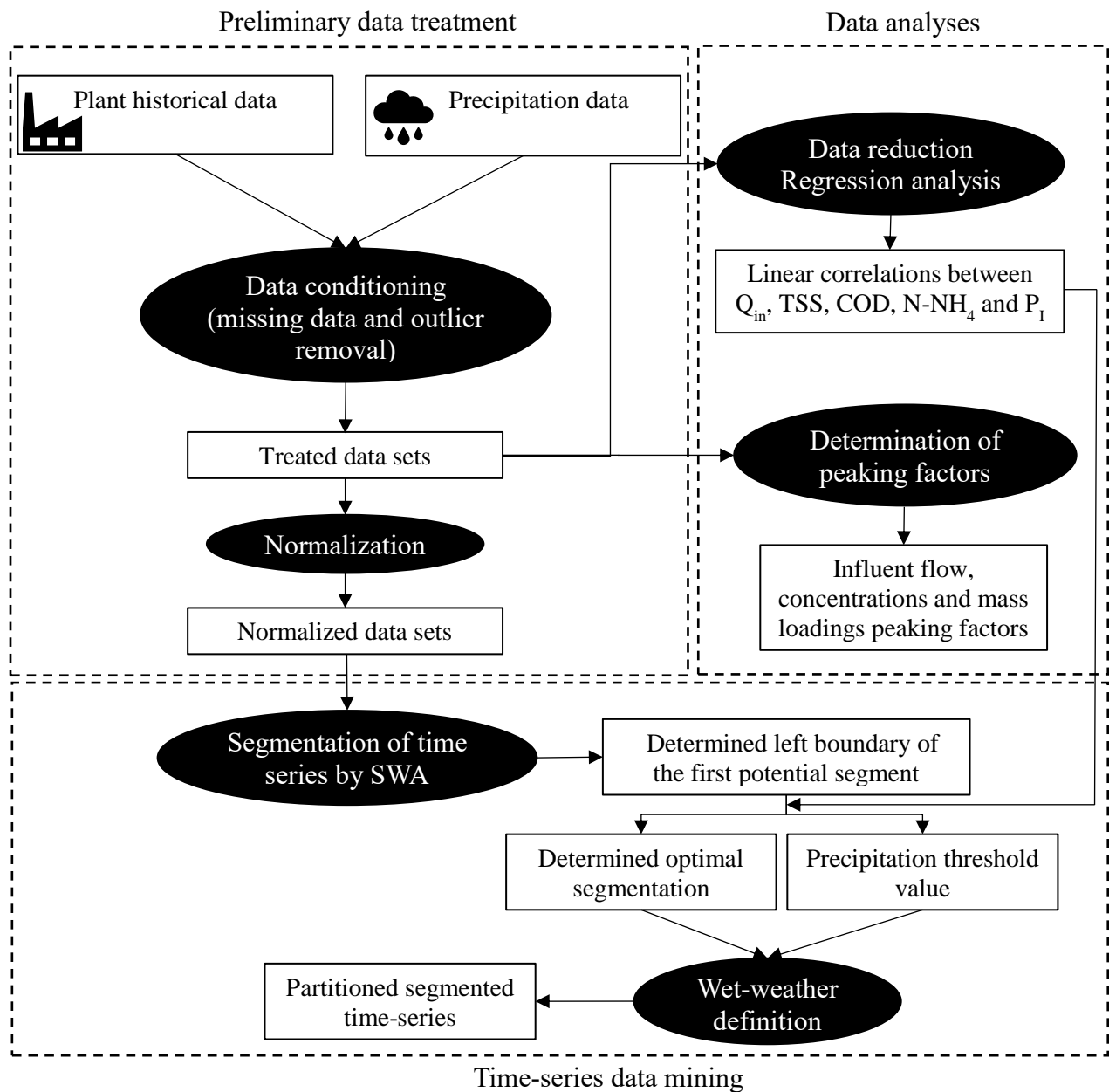


Fig. 3 A summary of the implemented methods in this study

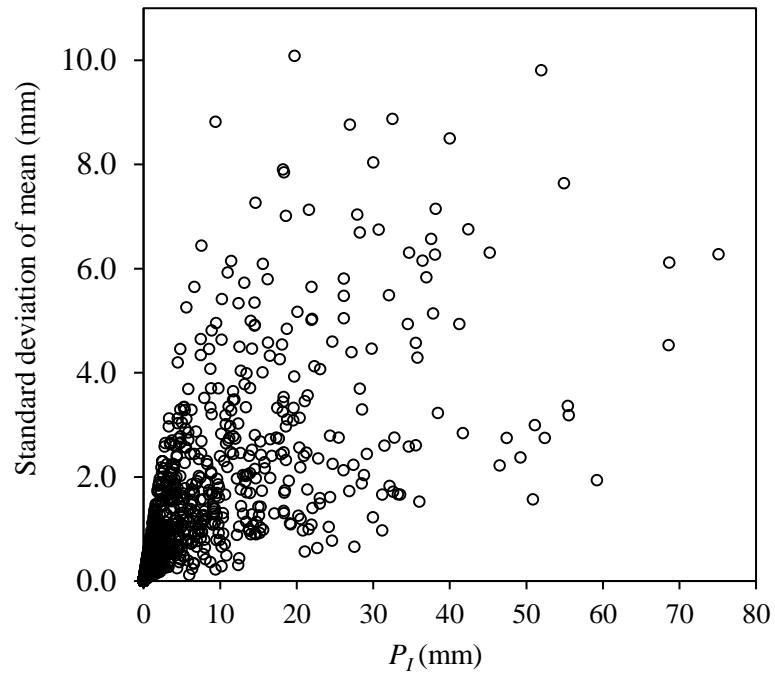


Fig. 4 Standard deviation of mean for P_I values

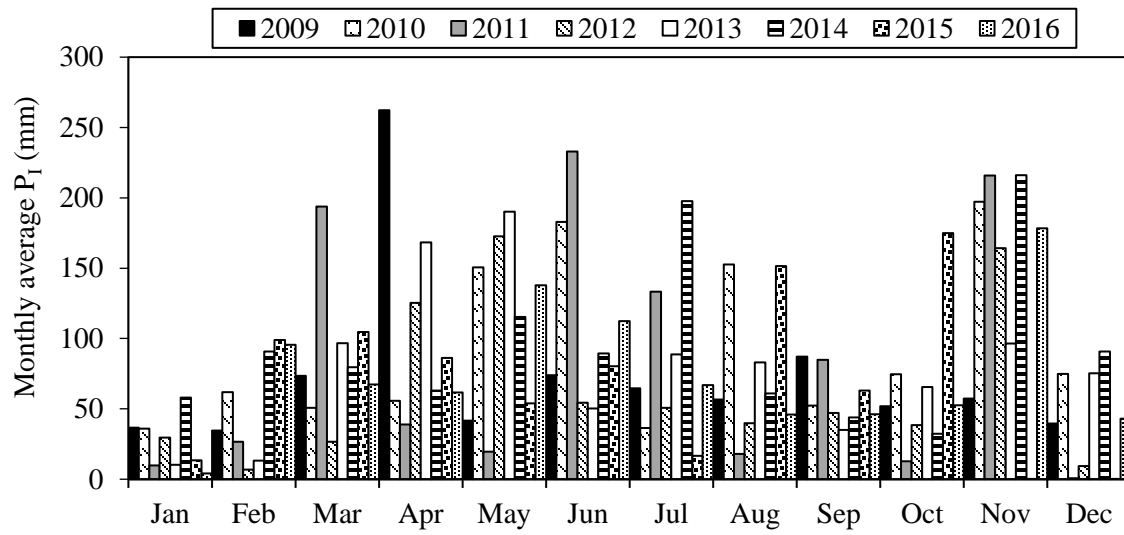


Fig. 5 Variation of monthly average P_I values during the studied period

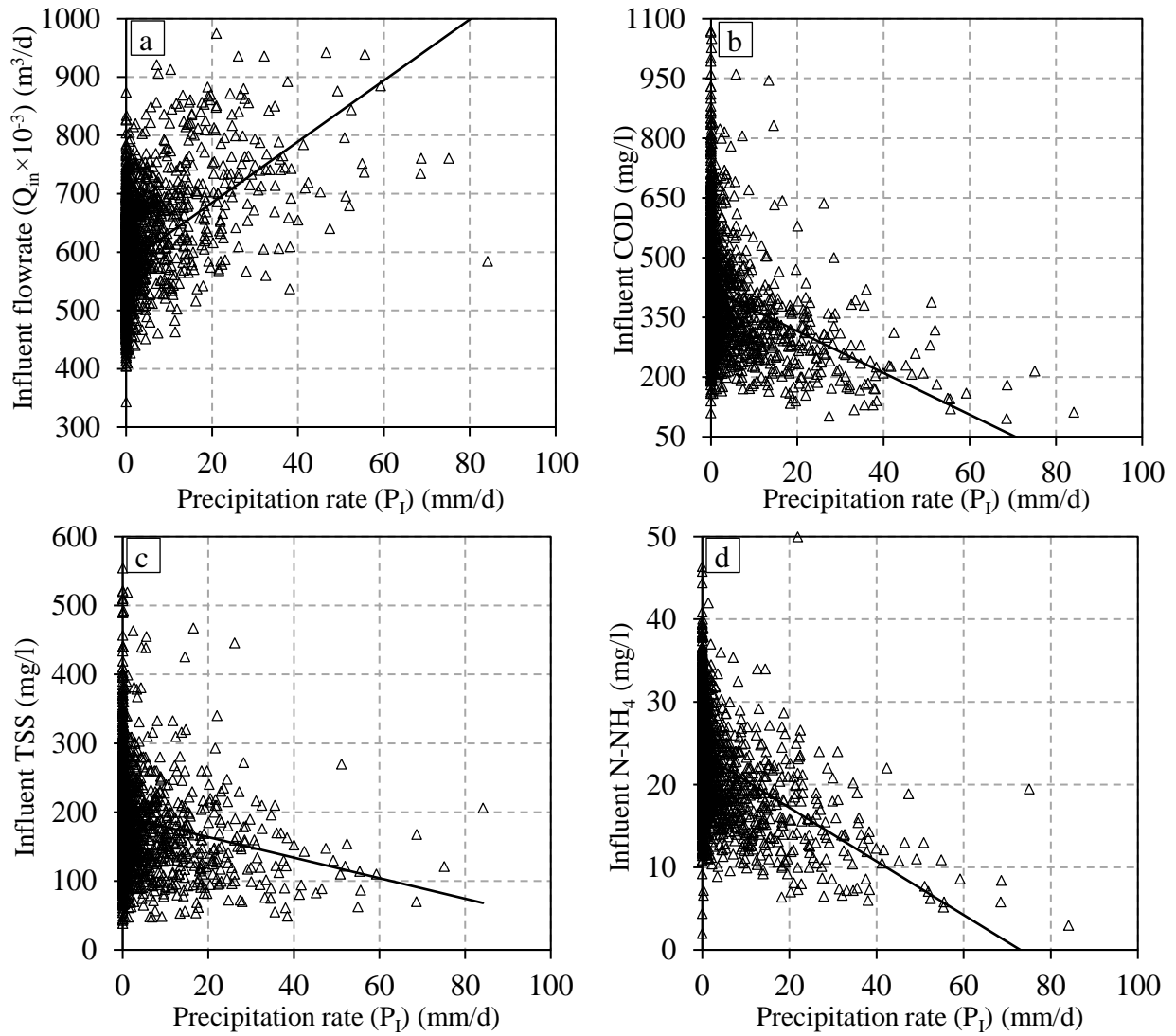


Fig. 6 Influent flow (a), TSS (b), COD (c) and $N-NH_4$ (d) versus precipitation intensity (P_1)

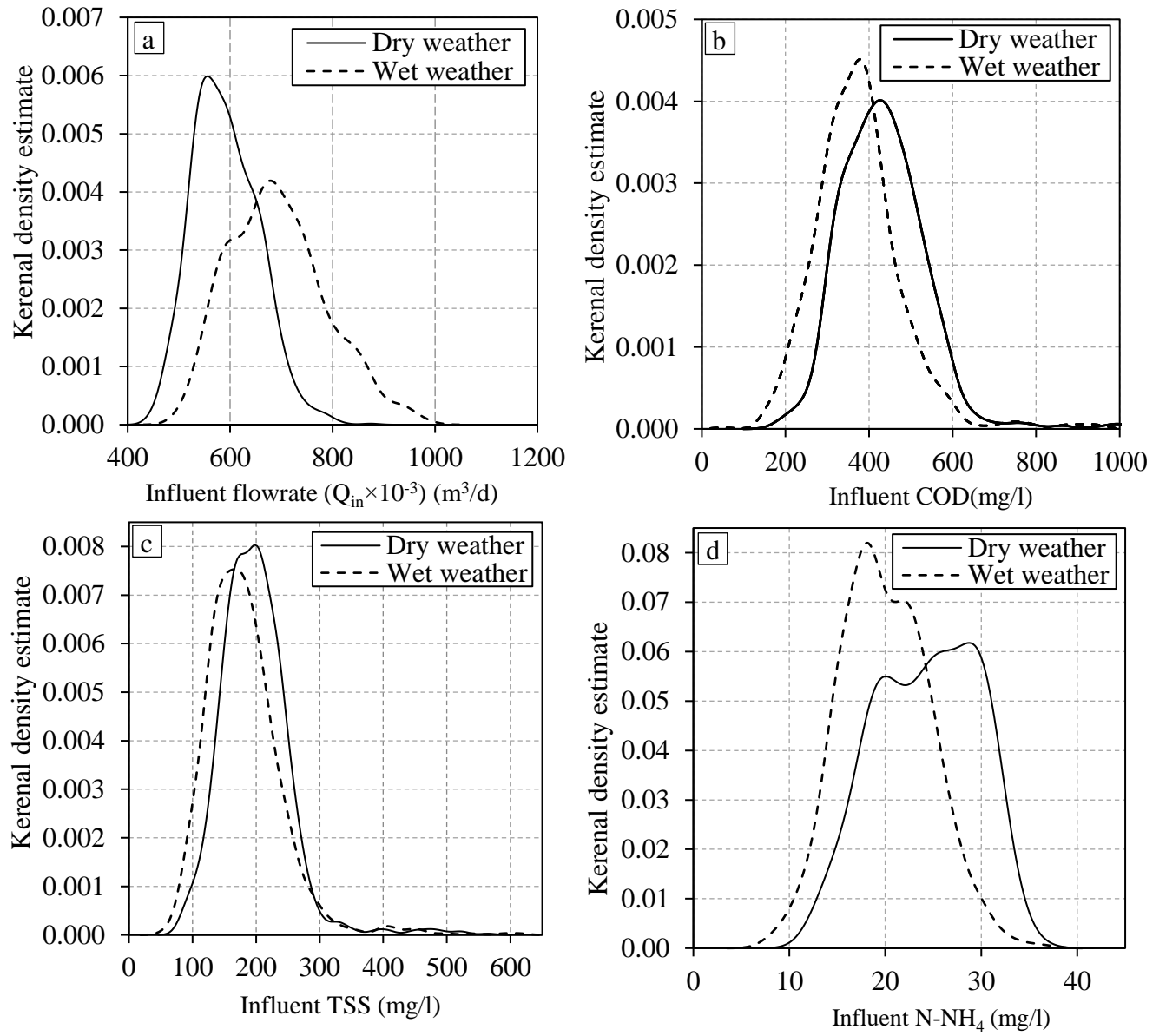


Fig. 7 Kernel density estimations of wet and dry weather influent flow (a), COD (b), TSS (c) and N-NH₄ (d)

Tables

Table 1. Basic statistics, number of missing and outlier values of total datasets

Table 2. Annual average of influent flowrate and concentrations

Table 3. Summary of influent flowrate, concentrations and loadings peaking factors

Table 4. Linear regression analysis and F-test results for entire data sets

Table 1. Basic statistics, number of missing and outlier values for datasets

Variable	Unit	Treated data		Outlier	Missing
		Mean	Standard deviation	Number	Number
TSS	mg. L ⁻¹	185	61.5	49	17
COD	mg. L ⁻¹	395	111.1	30	22
N-NH ₄	mg. L ⁻¹	22.24	5.98	23	15
Q _{in} [*]	m ³ . d ⁻¹	596.15	84.26	10	5
P _I	mm	2.75	7.36	-	21

*Results should be multiplied by 10³ to obtain the real data

Table 2. Annual average of influent flowrate and concentrations

Year		2009	2010	2011	2012	2013	2014	2015	2016
Variable	Unit								
TSS	mg. L ⁻¹	209.6	199.5	222.3	201.4	176.6	152.8	176.3	181.2
COD	mg. L ⁻¹	412.9	423.1	429.4	435.1	403.2	359.5	394.7	392.6
N-NH ₄	mg. L ⁻¹	23.7	23.7	22.9	22.4	22.8	20.6	21.4	25.7
Q _{in} [*]	m ³ . d ⁻¹	632.2	623.4	604.7	583.9	603.8	600.7	542.0	578.3

*Results should be multiplied by 10³ to obtain the real data

Table 3. Summary of influent flowrate, concentrations and loadings peaking factors

Parameter	Peaking factor	Formula	Average	Range
Flowrate (Q _{in})	P _{f1}	Q _{MMAD} : Q _{AD}	1.15	1.11-1.22
	P _{f2}	Q _{PD} : Q _{AD}	1.25	0.9-1.54
TSS	P _{C1}	C _{MMAD} : C _{AD}	1.36	1.32-1.70
	P _{C2}	C _{PD} : C _{AD}	1.77	1.41-3.33
	P _{L1}	L _{MMAD} : L _{AD}	1.34	1.18-1.55
	P _{L2}	L _{PD} : L _{AD}	2.24	1.44-4.38
COD	P _{C3}	C _{MMAD} : C _{AD}	1.26	1.22-1.32
	P _{C4}	C _{PD} : C _{AD}	1.65	1.25-2.62
	P _{L3}	L _{MMAD} : L _{AD}	1.54	1.12-3.61
	P _{L4}	L _{PD} : L _{AD}	1.81	0.95-5.34

C= COD and TSS concentrations

L= COD and TSS mass loadings

Table 4. Linear regression analysis and F-test results for entire data sets

Equation	R^2	F_{st}	F_{cr}
$Q_{in} = 5239.1 (P_I) + 579440$	0.25	0.008	0.93
$TSS = -1.41 (P_I) + 192.96$	0.03	0.014	0.93
$COD = -5.27 (P_I) + 418.81$	0.1	0.004	0.93
$N-NH_4 = -0.32 (P_I) + 23.7$	0.16	1.05	1.1

*Results should be multiplied by 10^3 to obtain the real data